Bayesian Nonparametric ROC Regression Modeling

Bayesian Analysis (2013) **8**, Number 3, pp. 623–646

Bayesian Nonparametric ROC Regression Modeling

Vanda Inácio de Carvalho, *† Alejandro Jara, ‡ Timothy E. Hanson, § and Miguel de Carvalho ¶‖

Abstract. The receiver operating characteristic (ROC) curve is the most widely used measure for evaluating the discriminatory performance of a continuous biomarker. Incorporating covariates in the analysis can potentially enhance information gathered from the biomarker, as its discriminatory ability may depend on these. In this paper we propose a dependent Bayesian nonparametric model for conditional ROC estimation. Our model is based on dependent Dirichlet processes, where the covariate-dependent ROC curves are indirectly modeled using probability models for related probability distributions in the diseased and healthy groups. Our approach allows for the entire distribution in each group to change as a function of the covariates, provides exact posterior inference up to a Monte Carlo error, and can easily accommodate multiple continuous and categorical predictors. Simulation results suggest that, regarding the mean squared error, our approach performs better than its competitors for small sample sizes and nonlinear scenarios. The proposed model is applied to data concerning diagnosis of diabetes.

Keywords: Conditional area under the curve, related probability distributions, dependent Dirichlet process, Markov chain Monte Carlo

1 Introduction

The statistical evaluation of diagnostic and screening procedures, such as biomarkers and imaging technologies, is of great importance in public health and medical research. The receiver operating characteristic (ROC) curve is a popular tool for evaluating the performance of continuous markers and it is widely used in medical studies. The ROC curve is a plot of the true positive rate (TPR; the probability that a diseased subject has a positive test) versus the false positive rate (FPR; the probability that a healthy subject has a positive test), across all possible threshold values used to classify subjects as healthy or diseased. That is, the ROC curve represents the plot $\{(\text{FPR}(k), \text{TPR}(k)) = (1 - F_0(k), 1 - F_1(k)), -\infty < k < \infty\}$, where F_0 and F_1 are the cumulative distribution functions of the marker in the healthy and diseased populations, respectively. For

*Department of Statistics, Pontificia Universidad Católica de Chile, Santiago, Chile, vanda.kinets@gmail.com
†Center for Statistics and Applications, University of Lisbon, Portugal
‡Department of Statistics, Pontificia Universidad Católica de Chile, Santiago, Chile, ajara@mat.puc.cl
§Department of Statistics, University of South Carolina, Columbia, US, hansont@stat.sc.edu
¶Department of Statistics, Pontificia Universidad Católica de Chile, Santiago, Chile, mdecarvalho@mat.puc.cl
‖Center for Mathematics and Applications, Universidade Nova de Lisboa, Portugal

 DOI:10.1214/13-BA825

$0 \leq u \leq 1$, the ROC curve is given by $\text{ROC}(u) = 1 - F_1\{F_0^{-1}(1-u)\}$. Related to the ROC curve, several measures, such as the area under the curve (AUC) or the Youden index, are considered as summaries of the discriminatory accuracy of the biomarker. The AUC, given by $\int_0^1 \text{ROC}(u)\text{d}u$, is most common—it is related to the Mann–Whitney statistic—and can be interpreted as the probability that the marker value of a diseased individual exceeds the one of a nondiseased individual.

It has been recently recognized that several factors can affect the marker distribution beyond the disease status (see for instance Pepe 1998, Faraggi 2003, González-Manteiga et al. 2011, Rodríguez-Álvarez et al. 2011a); examples of such factors include different test settings and subject-specific characteristics (Pepe 2003, Chapter 3). For instance, we are interested in evaluating the influence of age on the performance of blood glucose to accurately diagnose individuals with diabetes. It is therefore important to understand the influence of the covariates to determine the optimal and suboptimal conditions or populations to perform such tests on. Ignoring the covariate information may yield biased or oversimplified inferences, whereas stratifying by covariates may be either impractical (for continuous covariates) or incur a loss in power.

Several methods have been proposed to assess covariate effects on the ROC curve. The so-called "induced methodology" models the distribution of the marker in healthy and diseased populations separately and then computes the induced ROC curve (Pepe 1998; Faraggi 2003; González-Manteiga et al. 2011; Rodríguez-Álvarez et al. 2011a). Alternatively, direct methodology regresses the shape of the ROC curve directly onto covariates through a generalized linear model (Alonzo and Pepe 2002; Pepe 2003; Cai 2004). We refer the reader to Rodríguez-Álvarez et al. (2011b) for a comparative study of both methodologies. A crucial aspect of such methodologies is the use of parametric assumptions to model the effect of covariates on the ROC curve. For instance, the use of a parametric location model for F_0 and F_1 under the induced methodology framework may lead to misleading results if the effects are incorrectly specified. Although there is a vast literature dealing with nonparametric approaches for the estimation of ROC curves in the absence of covariates (Hsieh and Turnbull 1996; Zou et al. 1997; Lloyd 1998; Zhou and Harezlak 2002; Peng and Zhou 2004), few approaches have been developed for nonparametric estimation of the conditional ROC curve, in the presence of covariates.

The current approaches to ROC regression, within the induced context, are based on homoscedastic linear models with parametric errors (Faraggi 2003), parametric location models with unspecified error distributions (Pepe 1998), and heterocedastic nonparametric models based on kernel–type regression methods (see for instance González-Manteiga et al. 2011, Rodríguez-Álvarez et al. 2011a). In this work we propose a Bayesian nonparametric approach for modeling conditional ROC curves, within the induced methodology context. Our approach is based on dependent Dirichlet processes, and thus allows for the entire distribution to smoothly change as a function of covariates in the healthy and diseased groups and—unlike the kernel-based approaches—it allows for the inclusion of continuous and discrete predictors. Full inference for the covariate-specific ROC curves, as well as for the AUC, is easily obtained using Markov chain Monte Carlo (MCMC). Bayesian nonparametric techniques allow for broadening

the class of models under consideration and hence for the development of a widely applicable approach that can be used for practically any population and for a large number of diseases. Recent applications of Bayesian nonparametric models in ROC analysis can be found in Erkanli et al. (2006), Branscum et al. (2008), Hanson et al. (2008a), Hanson et al. (2008b), and Inácio et al. (2011).

The paper is organized as follows. Our modeling framework for the estimation of conditional ROC curves and its theoretical justification are presented in Section 2. In Section 3, a simulation study is carried out to assess and illustrate the performance of our model. In Section 4, the proposed methodology is applied to the analysis of diabetes data. Concluding remarks are given in Section 5.

2 The Bayesian nonparametric model

2.1 The modeling approach and its justification

Let y_{0i} and y_{1j} be real-valued continuous random variables denoting the marker result for the ith and jth subjects in the healthy and diseased group, respectively, $i = 1, \ldots, n_0$, $j = 1, \ldots, n_1$. Assume that p-dimensional covariate vectors $\mathbf{x}_{0i} \in \mathcal{X} \subset \mathbb{R}^p$ and $\mathbf{x}_{1j} \in \mathcal{X} \subset \mathbb{R}^p$ are recorded for the ith and jth subject in the healthy and diseased group, respectively. We assume that, given the covariates, the marker results are independent in the healthy and diseased groups and that

$$y_{0i} \mid \mathbf{x}_{0i} \overset{\text{ind.}}{\sim} f_0(\cdot \mid \mathbf{x}_{0i}), \quad i = 1, \ldots, n_0,$$

and

$$y_{1j} \mid \mathbf{x}_{1j} \overset{\text{ind.}}{\sim} f_1(\cdot \mid \mathbf{x}_{1j}), \quad j = 1, \ldots, n_1,$$

where $f_0(\cdot \mid \mathbf{x})$ and $f_1(\cdot \mid \mathbf{x})$ denote the conditional densities of the marker, given the predictors $\mathbf{x}$, in the healthy and diseased group, respectively.

We propose a model for the conditional ROC curves based on the specification of a probability model for the entire collection of densities $\mathcal{F}_0 = \{f_0(\cdot \mid \mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ and $\mathcal{F}_1 = \{f_1(\cdot \mid \mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$. Instead of specifying a Gaussian location–scale regression model for the marker values in each population (Pepe 1998; González-Manteiga et al. 2011; Rodríguez-Álvarez et al. 2011a), we model the conditional densities in each group using predictor-dependent mixtures of Gaussian models,

$$f_h(\cdot \mid \mathbf{x}) = \int \phi(\cdot \mid \mu, \sigma^2) \mathrm{d}G_{h\mathbf{x}}(\mu, \sigma^2), \quad h \in \{0, 1\},$$

where $\phi(\cdot \mid \mu, \sigma^2)$ denotes the density of the Gaussian distribution with mean μ and variance σ^2, and, for every $\mathbf{x} \in \mathcal{X}$, $G_{0\mathbf{x}}$ and $G_{1\mathbf{x}}$ are probability measures defined on $\mathbb{R} \times \mathbb{R}^+$. The probability model for the conditional densities is induced by specifying a probability model for the collection of mixing distributions $\mathcal{G}_0^{\mathcal{X}} = \{G_{0\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$ and $\mathcal{G}_1^{\mathcal{X}} = \{G_{1\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$. Justified by results in Barrientos et al. (2012), on the full support of models for predictor-dependent probability measures, we focused on

predictor-dependent discrete mixing distributions where only the support points are indexed by the predictor values. Specifically, we consider independent 'single–weights' dependent Dirichlet process priors (DDP) for $\mathcal{G}_0^{\mathcal{X}}$ and $\mathcal{G}_1^{\mathcal{X}}$ (MacEachern 2000). A 'single–weights' DDP prior involves a countable mixture of stochastic processes over $\mathcal{X}$, with weights matching those from the standard Dirichlet process (DP). Therefore, the prior for $\mathcal{G}_h^{\mathcal{X}}$, $h \in \{0,1\}$, has an almost sure discrete representation which extends the DP stick–breaking representation (Sethuraman 1994), where, for every $\mathbf{x} \in \mathcal{X}$,

$$G_{h\mathbf{x}}(\cdot) = \sum_{k=1}^{\infty} w_k^h \delta_{\boldsymbol{\theta}_k^h(\mathbf{x})}(\cdot), \quad h \in \{0,1\},$$

where $\delta(\cdot)$ denotes the Dirac measure, $\{\boldsymbol{\theta}_k^h(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$, $k \in \mathbb{N}$ and $h \in \{0,1\}$, are independent stochastic processes with index set $\mathcal{X}$, and the weights arise from a stick–breaking construction: $w_1^h = v_1^h$ and $w_k^h = v_k^h \prod_{r=1}^{k-1}(1 - v_r^h)$, for $k = 2, 3, \ldots$, with $v_r^h \mid \alpha_h \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha_h)$, for $\alpha_h \in \mathbb{R}^+$, independent across h and of the support point processes.

In our context, we consider $\boldsymbol{\theta}_k^h(\mathbf{x}) = (m_k^h(\mathbf{x}), \tau_k^h)' \in \mathbb{R} \times \mathbb{R}^+$, where $\{m_k^h(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$, $k = 1, 2, \ldots$, are i.i.d. Gaussian processes, with parameters $\boldsymbol{\Psi}_h$, and independent across h. The notation $\mathcal{G}_h^{\mathcal{X}} \mid \alpha_h, \boldsymbol{\Psi}_h \sim \text{DDP}(\alpha_h, \boldsymbol{\Psi}_h)$, $h \in \{0,1\}$, is used to denote the resulting DDP prior for the corresponding collection of predictor–dependent mixing distributions. Under this formulation, the resulting model for the conditional densities takes the form of an infinite mixture model

$$\begin{aligned} f_h(\cdot \mid \mathbf{x}) &= \int \phi(\cdot \mid \mu, \sigma^2) \mathrm{d}G_{h\mathbf{x}}(\mu, \sigma^2), \\ &= \sum_{k=1}^{\infty} w_k^h \phi(\cdot \mid m_k^h(\mathbf{x}), \tau_k^h), \quad h \in \{0,1\}. \end{aligned} \tag{1}$$

The conditional cumulative distributions can be expressed as

$$F_h(\cdot \mid \mathbf{x}) = \sum_{k=1}^{\infty} w_k^h \Phi(\cdot \mid m_k^h(\mathbf{x}), \tau_k^h), \quad h \in \{0,1\}.$$

Thus, for a given value of the covariate, the conditional ROC curve is defined, for $0 \leqslant u \leqslant 1$, as

$$\text{ROC}(u \mid \mathbf{x}) = 1 - F_1\{F_0^{-1}(1 - u \mid \mathbf{x}) \mid \mathbf{x}\}.$$

The use of our modeling approach for conditional ROC curves is justified by the full support property of the induced model under a product space topology, defined using the uniform norm. The following theorem is proved in Appendix A of the supplementary material.

Theorem 5. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be the underlying probability space associated with the DDP mixture of Gaussian distributions, with trajectories given by expression (1). For almost every $\omega \in \Omega$ and every $\mathbf{x} \in \mathcal{X}$, let $\text{ROC}_\omega(\cdot \mid \mathbf{x})$ be a trajectory of the ROC curve under*

the proposed DDP mixture model. Then, for every $T \in \mathbb{N}$, $\epsilon > 0$, *and* $\mathbf{x}_1, \ldots, \mathbf{x}_T \in \mathcal{X}$, *it follows that*

$$\mathbb{P}\left(\omega \in \Omega : \sup_{u \in [0,1]} |\mathrm{ROC}_\omega(u \mid \mathbf{x}_t) - \mathrm{ROC}(u \mid \mathbf{x}_t)| < \epsilon, t = 1, \ldots, T\right) > 0,$$

for every collection of continuous ROC curves $\{\mathrm{ROC}(\cdot \mid \mathbf{x}_t) : t = 1, \ldots, T\}$.

Theorem 1 establishes that the probability measure, induced by the use of fully specified DDP mixture models where only the support points are indexed by the predictors, assigns positive mass around any collection of ROC curves. Since alternative modeling frameworks could be considered, where only the weights or both weights and support points are indexed by predictors, the results summarized in Theorem 1 justify our modeling choice.

2.2 The B–splines DDP mixture model

Although flexible, priors such as the ones discussed in the previous section require sampling realizations of the Gaussian processes at each distinct value of the covariates and, thus, inferences could take prohibitively long to obtain. Therefore, we elaborate on a linear DDP (LDDP) prior formulation (De Iorio et al. 2004, 2009). Since the full support property of our proposal depends on the flexibility of the Gaussian processes defining the support points, we explore an approximation to the full model where the Gaussian processes are replaced by 'sufficiently rich' linear (in the coefficients) functions, $m_k^h(\mathbf{x}) = \mathbf{z}'\boldsymbol{\beta}_k^h$, where $\mathbf{z}$ is a q-dimensional design vector possibly including non-linear transformations of the continuous predictors. To this end, we consider an additive model formulation based on B–splines (see, e.g., Eilers and Marx 1996), referred to as B–splines DDP,

$$m_k^h(\mathbf{x}) = \beta_{k0}^h + \sum_{l=1}^{p} \left(\sum_{n=1}^{K_l} \beta_{kln}^h \psi_n(x_l, d_l) \right),$$

where $\psi_n(x, d)$ corresponds to the nth B–spline basis function of degree d, evaluated at x and $\boldsymbol{\beta}_k^h = \{\beta_{k0}^h, \ldots, \beta_{kpK_p}^h\}$. The previous formulation allows for the inclusion of discrete and continuous predictors.

It is important to stress that the theoretical result on the support of the process applies for a fully specified DDP mixture model, by considering well-defined Gaussian processes for the support point functions. The proposed B–splines DDP mixture model corresponds to an approximation of the fully specified DDP mixture model, where the well-defined Gaussian processes are approximated by finite–dimensional B–spline regressions; a standard practice in the nonparametric regression literature (see, e.g., Eilers and Marx 1996) and which is typically justified from a theoretical point of view by making assumptions on the smoothness of the functions to be approximated. Instead of providing a theoretical justification of the proposed model, by making unverifiable assumptions on the smoothness of the collections of 'true' densities and by assuming that the number of nodes goes to infinity, a simulation study is performed in Section

3 to illustrate the performance of the model under complex 'true' scenarios even when the number of nodes is very small ($K_l = 3$ is used in our applications).

Under the LDDP formulation, the base stochastic processes are replaced with a group-specific distribution G^*_{h0} that generates the component specific regression coefficients and variances. Therefore, the B–splines DDP mixture model can be equivalently formulated as a DP mixture of Gaussian regression models

$$f_h(\cdot \mid \mathbf{x}) \quad = \quad \int \phi(\cdot \mid \mathbf{z}'\boldsymbol{\beta}, \sigma^2) \mathrm{d}G_h(\boldsymbol{\beta}, \sigma^2), \tag{2}$$

and

$$G_h \mid \alpha_h, G^*_{0h} \overset{\text{ind.}}{\sim} \mathrm{DP}(\alpha_h, G^*_{0h}), \tag{3}$$

$h \in \{0, 1\}$. For each group, we consider normal–inverse–gamma distributions for the independent DP baselines, i.e.,

$$G^*_{0h} \equiv N_q(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h) \times \Gamma^{-1}(\tau_{h1}/2, \tau_{h2}/2), \tag{4}$$

where $N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the q-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\Gamma^{-1}(a, b)$ refers to the inverse–gamma distribution with parameters a and b. The model specification is completed by assuming, for $h \in \{0, 1\}$, the following independent hyper–priors:

$$\alpha_h \mid a_h, b_h \sim \Gamma(a_h, b_h), \quad \tau_{h2} \mid \tau_{s_{h1}}, \tau_{s_{h2}} \sim \Gamma(\tau_{s_{h1}}/2, \tau_{s_{h2}}/2), \tag{5}$$

$$\boldsymbol{\mu}_h \mid \mathbf{m}_h, \mathbf{S}_h \sim N_q(\mathbf{m}_h, \mathbf{S}_h), \quad \boldsymbol{\Sigma}_h \mid \nu_h, \boldsymbol{\Psi}_h \sim \mathrm{IW}_q(\nu_h, \boldsymbol{\Psi}_h), \tag{6}$$

where $\Gamma(a, b)$ refers to a gamma distribution with parameters a and b, and $\mathrm{IW}_q(\nu, \boldsymbol{\Psi})$ denotes a q-dimensional inverted–Wishart distribution with degrees of freedom ν and scale matrix $\boldsymbol{\Psi}$, parameterized such that $E(\boldsymbol{\Sigma}) = \boldsymbol{\Psi}^{-1}/(\nu - q - 1)$.

2.3 The prior specification

Many authors advocating infinite mixture models choose hyperprior values that seem reasonable, and in fact are reasonable for the data they consider (De Iorio et al. 2009; Jara et al. 2010). Here, following the literature on finite mixture models (see, e.g., Richardson and Green 1997; Xu et al. 2010), we develop reasonable data-driven priors that encourage mixture components within a certain size range and complexity. We emphasize that the prior is data-driven mostly regarding the predictors, which are treated as fixed, and not the response.

Let $\tau = \sigma^{-2}$ be a Gaussian precision parameter and assume that $\tau \mid a, b \sim \Gamma(a/2, b/2)$ and $b \sim \Gamma(c/2, d/2)$. It follows that the marginal distribution for τ is a compound gamma distribution (Dubey 1970),

$$p(\tau) = \int_0^\infty f(\tau \mid b) f(b) \mathrm{d}b = \frac{\tau^{\frac{a}{2}-1} d^{\frac{c}{2}} \Gamma\left(\frac{a+c}{2}\right)}{\Gamma(\frac{a}{2})\Gamma(\frac{c}{2})(\tau + d)^{\frac{a+c}{2}}}, \quad \tau > 0.$$

This can be used to show that

$$E(\sigma^2) = \frac{c}{d(a-2)}, \quad \text{var}(\sigma^2) = \frac{2c(a+c-2)}{d^2(a-2)^2(a-4)}.$$

The improper prior $p(\tau) \propto 1/\tau$ is approximated by $a = 2, c = \epsilon, d = \epsilon$, where ϵ is small (e.g. $\epsilon = 0.001$). Clearly, $a > 2$ for the mean to exist and $a > 4$ for the variance to exist. If we have prior guesses $\eta = E(\sigma^2)$ and $v = \text{var}(\sigma^2)$, then solving the system of nonlinear equations for c and d yields

$$c = \frac{2\eta^2(2-a)}{2\eta^2 + 4v - av}, \quad d = \frac{-2\eta}{2\eta^2 + 4v - av}.$$

To keep $c > 0$ and $d > 0$, an a such that $a > 4 + 2\eta^2/v$ is required. An estimate $\hat{\sigma}^2$ from fitting a single trend, i.e. with $G_{h\mathbf{x}}(\cdot) = \delta_{\boldsymbol{\theta}^h(\mathbf{x})}(\cdot)$, serves as an upper bound, as mixtures of flexible regressions can only decrease variability. However, in some parts of the predictor space there may be essentially only one regression necessary, and so $\hat{\sigma}^2$ should be within the realm of non–negligible mass under the prior. One possible prior might be $\eta = v = \hat{\sigma}^2/4$. The rule-of-thumb that a random variable is within one standard deviation of its mean 68% of the time would imply that the value $\hat{\sigma}^2$ should be well–supported under the prior. We need $a > 4 + 0.5\hat{\sigma}^2$. Increasing a pushes mass toward zero and infinity for fixed mean and variance, i.e., gives weight to really big and/or really small precisions; setting $a = 5 + 0.5\hat{\sigma}^2$ seems reasonable. Collecting all of this together, in the context of our model, we propose to set

$$\tau_{h1} = 5 + 0.5\hat{\sigma_h}^2, \quad \tau_{s_{h1}} = \frac{(\tau_{h1} - 2)\hat{\sigma_h}^2}{2\tau_{h1} - 8 - \hat{\sigma_h}^2}, \quad \tau_{h2} = 2, \quad h \in \{0, 1\}.$$

We now turn attention to the prior specification for the mean and covariance matrix of the normal centering distribution. Using the same linear predictor as in the B–splines DDP mixture model, let $\widehat{\boldsymbol{\mu}}_h$ be the least squares estimate $\widehat{\boldsymbol{\mu}}_h = (\mathbf{Z}_h'\mathbf{Z}_h)^{-1}\mathbf{Z}_h'\mathbf{y}_h$, and let $\hat{\sigma}_h^2 = \|\mathbf{y}_h - \mathbf{Z}_h\widehat{\boldsymbol{\mu}}_h\|^2/(n_h - p)$, where $n_0 = n$ and $n_1 = m$. Under normal theory, $\boldsymbol{\mu}_h \sim N_q(\widehat{\boldsymbol{\mu}}_h, \hat{\sigma}_h^2(\mathbf{Z}_h'\mathbf{Z}_h)^{-1})$, approximately. Thus, we propose to set $\mathbf{m}_h = \widehat{\boldsymbol{\mu}}_h$ and $\mathbf{S}_h = \hat{\sigma}_h^2(\mathbf{Z}_h'\mathbf{Z}_h)^{-1}$. Finally, it might make sense to allow for more variability than $\hat{\sigma}_h^2(\mathbf{Z}_h'\mathbf{Z}_h)^{-1}$ for the inverted–Wishart prior distribution for $\boldsymbol{\Sigma}_h$, allowing for quite different regression coefficients in the mixture model. We considered 5 standard deviations, set $\nu = q + 2$ and $\boldsymbol{\Psi}_h^{-1} = 25\hat{\sigma}_h^2(\mathbf{Z}_h'\mathbf{Z}_h)^{-1}$.

2.4 Posterior inference

We use a marginal Gibbs sampling algorithm (MacEachern 1994; MacEachern and Müller 1998; Neal 2000) for simulation from the posterior distribution arising from expressions (2)–(6). Under this approach, the mixing distributions G_0 and G_1 are integrated out from the model and the algorithm uses the Polya urn representation of the DP predictive measure (Blackwell and MacQueen 1973). A full description of the conditional distributions needed for the implementation of the algorithm is given in Appendix B of the supplementary material.

Inferences on the induced conditional ROC curves require the sampling of the DP random measures. To this end, the ϵ–DP approximation (Muliere and Tardella 1998) is employed. Under this approach samples of the DP random measure are approximated by a finite-dimensional discrete distribution such that the total variation distance between the full realization and the approximation is smaller than ϵ; the value $\epsilon = 0.01$ is used in the applications of the model.

Finally, the computation of the induced conditional ROC curve requires the evaluation of the quantile function of a mixture of Gaussian distributions, which is computed numerically. The algorithm previously described is implemented in the function `LDDProc`, of the library DPpackage (Jara 2007; Jara et al. 2011), in the R program (R Development Core Team 2012).

3 A simulation study

To evaluate the performance of the estimators associated with our model, we analyzed simulated data sets under three different scenarios. Specifically, we considered a linear–mean scenario, a nonlinear–mean scenario with constant variance, and a nonlinear–mean scenario with predictor dependent variance and multimodality. For each of the scenarios 100 data sets were generated for each of the sample sizes: $n \equiv n_0 = n_1 = 50, 100, 200$. Using the simulated data sets the proposed model was compared with its main competitors. Given that the nonparametric kernel estimator can only handle univariate continuous predictors, we restricted the simulation study to this framework.

3.1 The simulation scenarios

In the first case (Scenario I), we consider different homoscedastic linear-mean regression models for the diseased and healthy groups. Specifically, we assume that, for $i = 1, \ldots, n$,

$$y_{0i} \mid x_{0i} \overset{\text{ind.}}{\sim} N\left(0.5 + x_{0i}, 1.5^2\right), \quad y_{1i} \mid x_{1i} \overset{\text{ind.}}{\sim} N\left(2 + 4x_{1i}, 2^2\right).$$

The purpose of including this linear scenario is to ascertain the loss of efficiency of the estimator associated to our model when the standard parametric assumptions hold.

In Scenario II, we assume different homoscedastic nonlinear-mean normal regression models for both groups:

$$y_{0i} \mid x_{0i} \overset{\text{ind.}}{\sim} N\left(\sin\{\pi \times (x_{0i} + 1)\}, 0.5^2\right), \quad y_{1i} \mid x_{1i} \overset{\text{ind.}}{\sim} N\left(0.5 + x_{1i}^2, 1^2\right).$$

Finally, in Scenario III we assume non-standard regression models for both groups. In this case, a two-component mixture of normals model with non-linear mean function and smoothly changing non-unimodal conditional distribution for the diseased group. For the healthy group, a heteroscedastic nonlinear-mean normal regression model was assumed. Specifically, we assume that, for $i = 1, \ldots, n$,

$$y_{0i} \mid x_{0i} \overset{\text{ind.}}{\sim} N\left(\sin(\pi x_{0i}), 0.2 + 0.5\exp(x_{0i})\right)$$

and

$$y_{1i} \mid x_{1i} \overset{\text{ind.}}{\sim} \frac{\exp(x_{1i})}{1+\exp(x_{1i})} N\left(x_{1i}, 0.5^2\right) + \frac{1}{1+\exp(x_{1i})} N\left(x_{1i}^3, 1^2\right).$$

In all cases, the predictor values were independently generated from a uniform distribution, $x_{0i} \overset{\text{i.i.d.}}{\sim} U(-1,1)$ and $x_{1i} \overset{\text{i.i.d.}}{\sim} U(-1,1)$.

3.2 The models

For each simulated dataset we fit the B–splines DDP mixture model by assuming $K_1 = 3$ and the prior specification described in Section 2.3. In all cases, 2000 MCMC samples were kept after a burn-in period of 2000 scans of the posterior distribution. Our model was compared with the semiparametric approach of Pepe (1998) and the nonparametric kernel estimator of Rodríguez-Álvarez et al. (2011a). González-Manteiga et al. (2011) and Rodríguez-Álvarez et al. (2011a) proposed nonparametric kernel estimators, whose main difference is the order of the local polynomial smoothers used for estimating the regression functions. González-Manteiga et al. (2011) employed a local constant fit (order 0), while Rodríguez-Álvarez et al. (2011a) considered a linear fit (order 1). Since local constant regression suffers from boundary–bias problems (see, e.g. Fan and Gijbels 1996), we only considered the approach of Rodríguez-Álvarez et al. (2011a). Furthermore, in addition to the original approach proposed by Pepe (1998), we considered an extension of this approach by using a B–splines trend. To distinguish between these semiparametric approaches, we have designated Pepe's original estimator as semiparametric linear and the other as semiparametric B–splines. For the implementation of the kernel estimator, regression and variance functions were estimated using local linear and local constant fits, respectively. The Gaussian kernel was chosen and generalized cross-validation was used to select the optimal bandwidth; more details on the implementation of the kernel-based approach is given in Appendix C of the supplementary material.

3.3 The results

Following Rodríguez-Álvarez et al. (2011a) and González-Manteiga et al. (2011), the discrepancy between estimated and true ROC curves was measured using the empirical global mean squared error

$$\begin{aligned}
\text{EGMSE} &= \frac{1}{n_x}\sum_{l=1}^{n_x}\frac{1}{n_u}\sum_{r=1}^{n_u}\left\{\widehat{\text{ROC}}(u_r \mid x_l) - \text{ROC}(u_r \mid x_l)\right\}^2 \\
&\approx \int_{\mathcal{X}}\int_0^1 \left\{\widehat{\text{ROC}}(u \mid x) - \text{ROC}(u \mid x)\right\}^2 \mathrm{d}u\mathrm{d}x \\
&= \mathbb{E}_{\mathcal{X}}\left[\int_0^1 \left\{\widehat{\text{ROC}}(u \mid x) - \text{ROC}(u \mid x)\right\}^2 \mathrm{d}u\right],
\end{aligned}$$

where $n_x = 25$, $n_u = 100$, and x_l and u_r lay on an evenly–spaced grid over the predictor space $\mathcal{X}$ and $[0,1]$, respectively. Table 1 and Figure 1 summarize the EGMSE for each

scenario, approach, and sample size.

Scenario	n	Approach			
		Sem. Linear	Sem. B–splines	Kernel	B–splines DDP
I	50	0.0084 (0.0057)	0.0140 (0.0080)	0.0131 (0.0073)	0.0138 (0.0075)
	100	0.0045 (0.0026)	0.0076 (0.0048)	0.0074 (0.0043)	0.0079 (0.0048)
	200	0.0022 (0.0014)	0.0037 (0.0023)	0.0036 (0.0020)	0.0042 (0.0022)
II	50	0.0385 (0.0056)	0.0122 (0.0058)	0.0130 (0.0064)	0.0125 (0.0061)
	100	0.0364 (0.0037)	0.0076 (0.0037)	0.0079 (0.0041)	0.0079 (0.0039)
	200	0.0345 (0.0022)	0.0045 (0.0015)	0.0042 (0.0017)	0.0047 (0.0017)
III	50	0.0534 (0.0090)	0.0218 (0.0112)	0.0302 (0.0156)	0.0162 (0.0090)
	100	0.0499 (0.0057)	0.0127 (0.0052)	0.0155 (0.0064)	0.0091 (0.0055)
	200	0.0470 (0.0036)	0.0091 (0.0032)	0.0098 (0.0041)	0.0062 (0.0031)

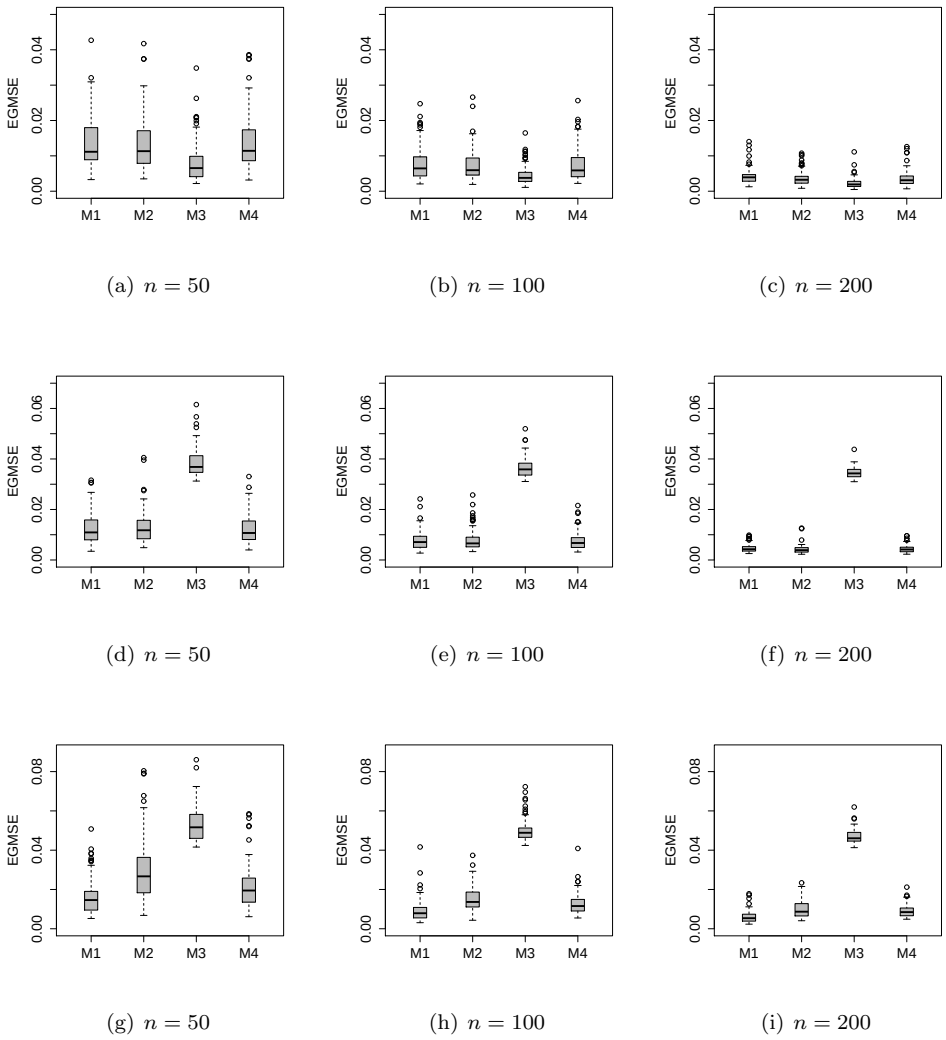
Table 1: Simulated data: Average (standard deviation), across simulations, of the empirical global mean squared error of the ROC curve for the different approaches under consideration. The results are presented for each of the simulation scenarios and sample sizes (n).

As expected, under the linear scenario (Scenario I), the semiparametric linear approach showed the best performance. The kernel, the semiparametric B–splines and the B–splines DDP mixture model have similar performances, although the kernel estimator was slightly better. The higher EGMSE values observed for the B–splines DDP mixture model, kernel and semiparametric B–splines estimators are explained by the bigger variability of the corresponding estimates when a simple parametric model holds. The difference between the semiparametric linear approach and the other competitors decreases as the sample size increases, which is explained by the reduction in the variance of the estimators for the more flexible models. Figure 2 depicts the estimated AUC function, along with the 2.5% and 97.5% simulation quantiles, under Scenario I. In this case, all methods recovered the functional form of the true AUC function successfully. Again, the semiparametric linear estimator showed a better performance than the other competitors and the difference between the estimators vanished as the sample size increases.

Under Scenario II, the results show that Pepe's semiparametric linear approach is clearly unsuitable and, as expected, its poor performance fails to improve as the sample size increases. Figure 3 shows that the B–splines DDP mixture model, the kernel method and the semiparametric B–splines estimator successfully recover the form of the true AUC function and illustrate that misleading results can be obtained using the semiparametric linear approach for this important functional.

Finally, under Scenario III, the B–splines DDP mixture model clearly outperformed the kernel method and the semiparametric B–splines approach for all sample sizes. Again the semiparametric linear approach showed poor behavior. Figure 4 shows that the nonparametric estimators recover the true AUC function successfully, whereas the semiparametric linear estimator produces misleading results.

The results of the simulation studies, therefore, strongly suggest that precise estimates of the conditional ROC curves and other functionals of interest can be obtained under the B–splines DDP mixture model.

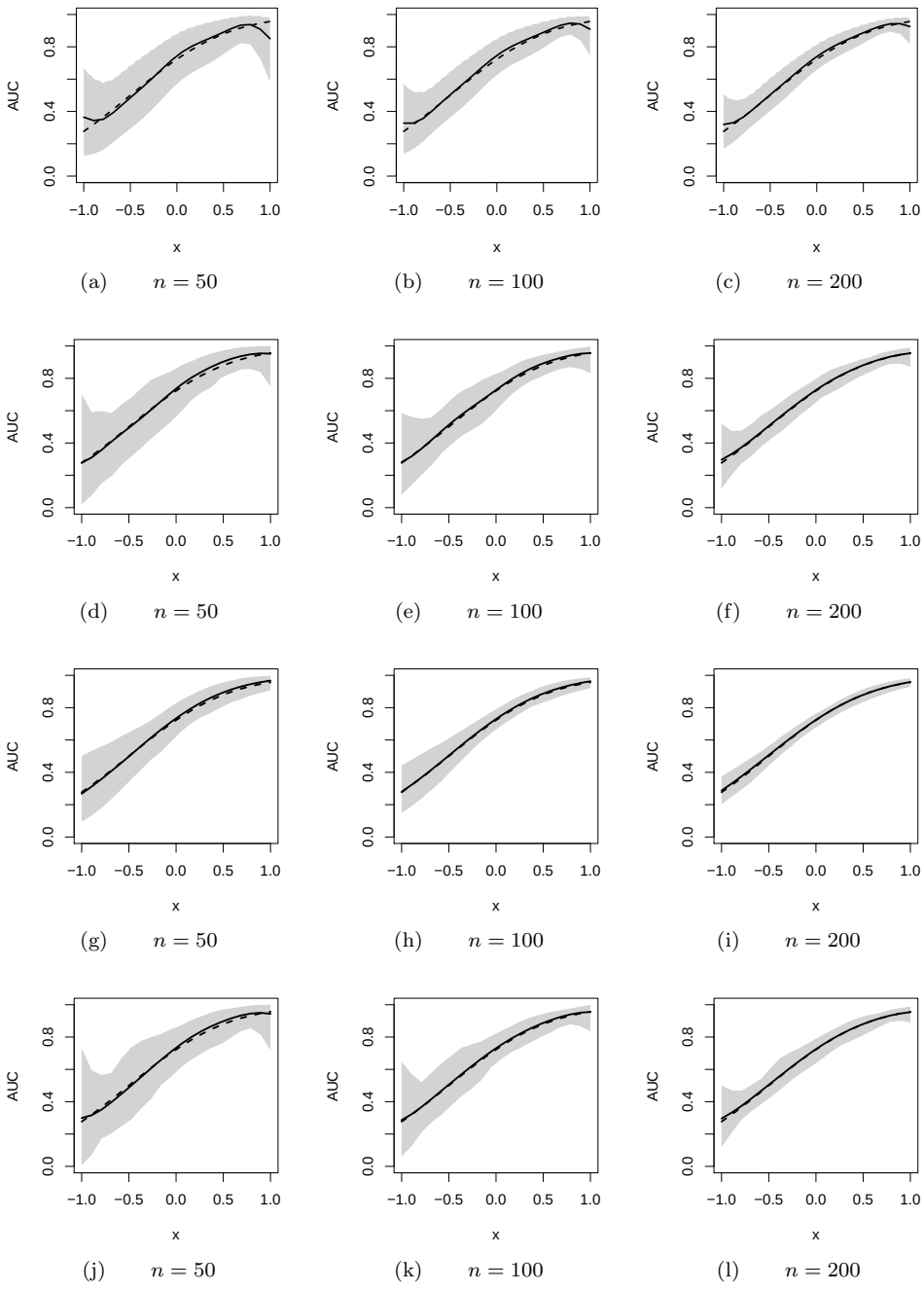


(a) $n = 50$ (b) $n = 100$ (c) $n = 200$

(d) $n = 50$ (e) $n = 100$ (f) $n = 200$

(g) $n = 50$ (h) $n = 100$ (i) $n = 200$

Figure 1: Simulated data: Box plots of the empirical global mean squared error (EGMSE) across simulations for the B–splines DDP mixture model (M1), kernel estimator (M2), semiparametric linear estimator (M3) and semiparametric B–splines estimator (M4). Panels (a)–(c), (d)–(f) and (g)–(i) and (j)–(l) display the results for Scenario I, II and III, respectively.



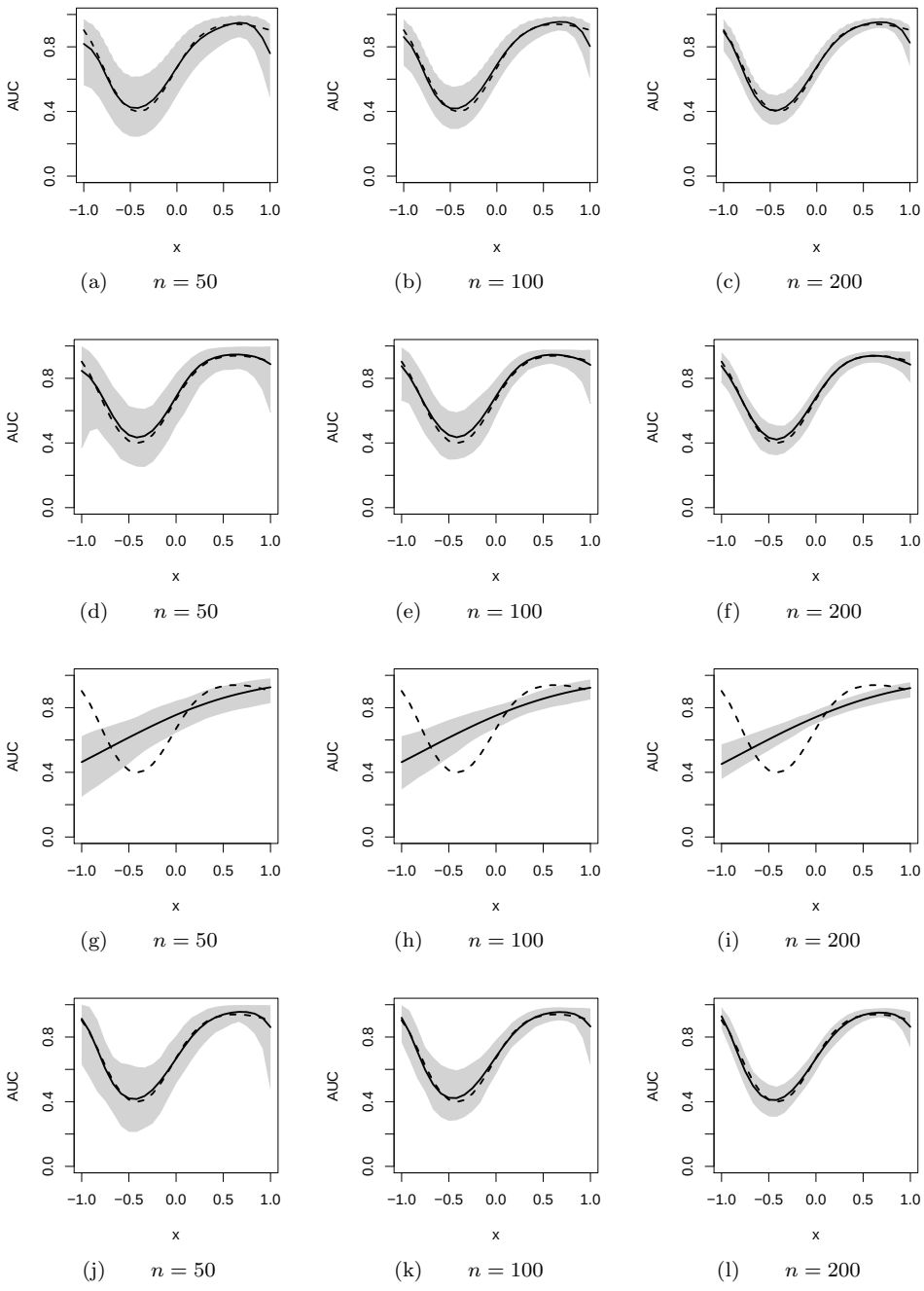
(a) $n = 50$ (b) $n = 100$ (c) $n = 200$

(d) $n = 50$ (e) $n = 100$ (f) $n = 200$

(g) $n = 50$ (h) $n = 100$ (i) $n = 200$

(j) $n = 50$ (k) $n = 100$ (l) $n = 200$

Figure 2: Simulated data: True (dotted line) and mean across simulations (solid line) of the posterior mean of the AUC function under Scenario I. A band constructed using the point-wise 2.5% and 97.5% quantiles across simulations is presented in gray. Panels (a)–(c), (d)–(f), (g)–(i) and (j)–(l) display the results for the B–splines LDDP mixture model, kernel, semiparametric linear and semiparametric B–splines approaches, respectively, for the sample sizes under consideration.



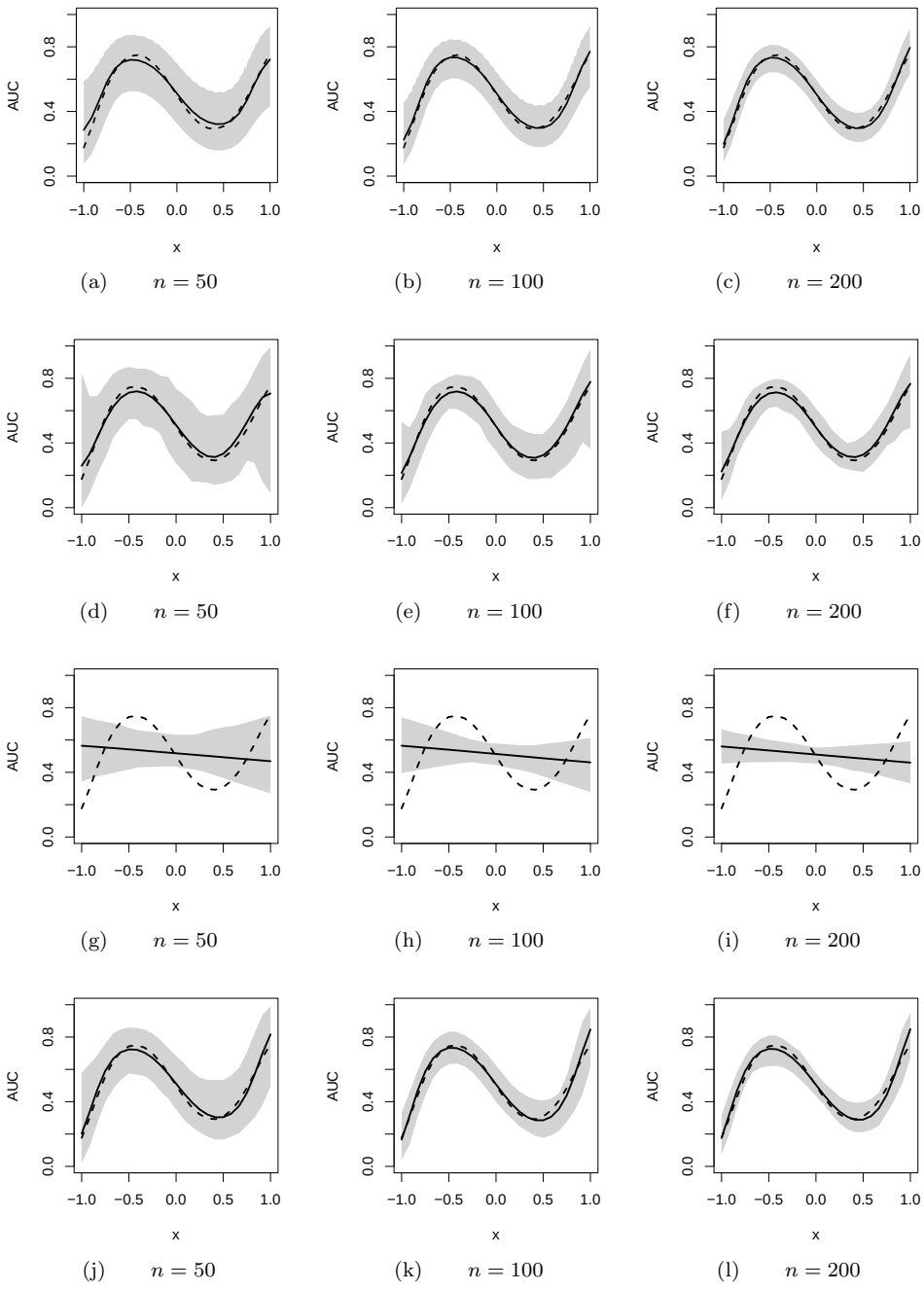
(a) $n = 50$ (b) $n = 100$ (c) $n = 200$

(d) $n = 50$ (e) $n = 100$ (f) $n = 200$

(g) $n = 50$ (h) $n = 100$ (i) $n = 200$

(j) $n = 50$ (k) $n = 100$ (l) $n = 200$

Figure 3: Simulated data: True (dotted line) and mean across simulations (solid line) of the posterior mean of the AUC function under Scenario II. A band constructed using the point-wise 2.5% and 97.5% quantiles across simulations is presented in gray. Panels (a)–(c), (d)–(f), (g)–(i) and (j)–(l) display the results for the B–splines DDP mixture model, kernel method, the semiparametric linear and semiparametric B–splines approaches, respectively, for the sample sizes under consideration.



(a) $n = 50$ (b) $n = 100$ (c) $n = 200$

(d) $n = 50$ (e) $n = 100$ (f) $n = 200$

(g) $n = 50$ (h) $n = 100$ (i) $n = 200$

(j) $n = 50$ (k) $n = 100$ (l) $n = 200$

Figure 4: Simulated data: True (dotted line) and mean across simulations (solid line) of the posterior mean of the AUC function under Scenario III. A band constructed using the point-wise 2.5% and 97.5% quantiles across simulations is presented in gray. Panels (a)–(c), (d)–(f), (g)–(i) and (j)–(l) display the results for the B–splines DDP mixture model, kernel method, semiparametric linear and semiparametric B–splines approaches, respectively, for the sample sizes under consideration.

3.4 The sensitivity analysis

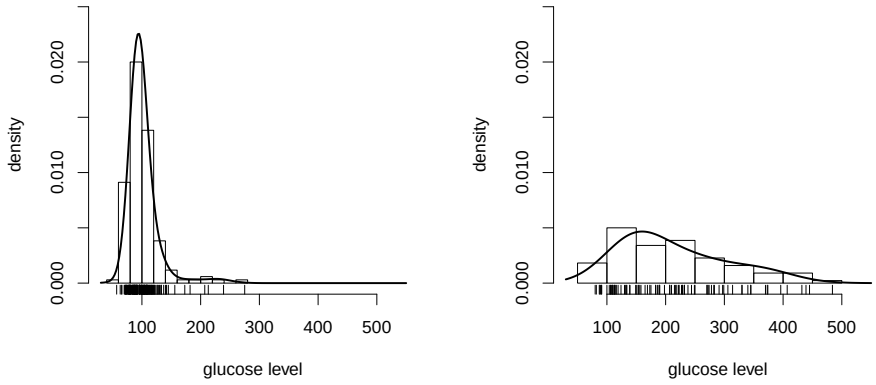
To investigate the influence of the specification of the hyper-parameter values we carried out a sensitivity analysis using the hyper-parameter values that are commonly used in the literature (De la Cruz et al. 2007; Jara et al. 2010, 2011). Specifically, we set: $a_0 = a_1 = 5$, $b_0 = b_1 = 1$, $\mathbf{m}_0 = \mathbf{m}_1 = (0, 0, 0, 0)$, $\mathbf{S}_0 = \mathbf{S}_1 = 10^2 \times \mathbf{I}_4$, $\nu_0 = \nu_1 = 6$, $\boldsymbol{\Psi}_0 = \boldsymbol{\Psi}_1 = \mathbf{I}_4$, $\tau_{s_{01}} = \tau_{s_{11}} = 6.01$, $\tau_{s_{02}} = \tau_{s_{12}} = 2.01$, and $\tau_{01} = \tau_{11} = 6.01$. To distinguish the resulting models under these prior specifications, in what follows, we have designated the B-splines DDP mixture model under these hyper-parameter values as B-splines DDP II. The results are shown in Appendix D of the supplementary material. As can be seen, the results of both prior specifications are the same.

4 Application to diabetes diagnosis

4.1 Data description

Diabetes is a metabolic disease mainly characterized by high blood sugar concentration and insulin deficiency or resistance. It is believed that the aging process may be associated with relative insulin deficiency or resistance among persons who are healthy (Smith and Thompson 1996). Diabetes doubles the risk of cardiovascular disease (Sarwar et al. 2010). In 2000, according to the World Health Organization (WHO), at least 171 million people worldwide suffered from diabetes, which corresponds to 2.8% of the World population. Its incidence is increasing rapidly, and it is estimated by 2030, this number will almost double (Wild. et al. 2004).

Our motivating data set comes from a population–based pilot survey of diabetes in Cairo, Egypt, in which postprandial blood glucose measurements were obtained from a fingerstick on 286 subjects. The gold standard for diagnosing diabetes, according to the WHO criteria, consists of a fasting plasma glucose value $\geqslant 140$ mg/dl or a 2 hour plasma glucose value $\geqslant 200$ mg/dl following a 75g oral glucose challenge (Smith and Thompson 1996). Based on these criteria 88 subjects were classified as diseased and 198 as healthy. These data have also been analyzed in Smith and Thompson (1996), Faraggi (2003), and in González-Manteiga et al. (2011). In the analyses presented here, we considered a subset of 258 subjects with age ranging from 27 to 78 years old. We restricted the analysis to this range of age because both groups had observations there. Figure 5 shows the histograms of the glucose levels for the healthy and diseased groups, along with DPM mixture of normals estimates of the densities. As expected, the distribution of glucose concentration in the diseased group tends to have more probability mass for higher values than the corresponding distribution of the healthy group. An initial analysis of the data, using independent DPM mixture of normal models for the log–concentration of glucose in both groups, showed a good marginal discriminatory performance of the marker to detect patients having a higher risk of diabetes. The estimated ROC curve from the preliminary analysis is shown in Figure 6. The corresponding estimated AUC is 0.885 $(0.823, 0.934)$.



(a) (b)

Figure 5: Glucose data: Histogram of the glucose concentration in the healthy (Panel a) and diseased group (Panel b). The posterior mean of the density for each group under (independent) DPM of normals models is displayed as a solid line.

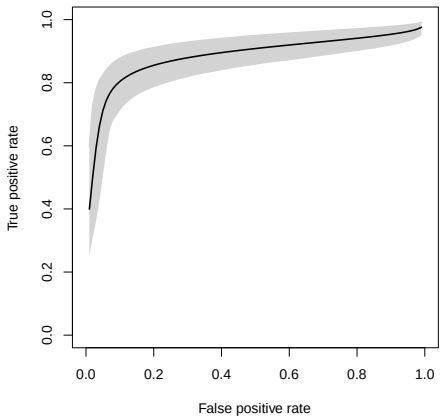


Figure 6: Diabetes data: Estimated ROC curve of the glucose levels (with no age effect). The estimate was obtained using a DPM mixture of normal models. The posterior mean (solid line) is presented along with the point-wise 95% highest posterior density (HPD) intervals.

4.2 The results

We fit the B–splines DDP mixture model for the glucose concentration by assuming $K_1 = 3$ and the prior specification described in Section 2.3. Figure 7 presents the estimated predictive density of glucose concentration on the healthy and diseased groups. The results suggest that the glucose level is much more concentrated in the healthy than

in the diseased group, across age. Interestingly, the healthy group showed a positive linear behavior in the conditional location of the glucose concentration and with an almost constant variability across age. On the other hand, the diseased group showed a nonlinear behavior on the location of the conditional distributions and a reduction of the dispersion as the age increases. The linear and nonlinear behavior of the location of the conditional distributions in the healthy and diseased group, respectively, is illustrated in panels (g) and (f) of Figure 7, which shows the posterior inference for the conditional mean functions. In the healthy population, the older the subject the higher the glucose level, which is in agreement with the results by Smith and Thompson (1996), who suggested that the aging process is associated with relative insulin deficiency or resistance among people who are healthy.

The posterior mean of the conditional ROC curves across age is shown in Figure 8 (a). In Figure 8 (c), (d), (e), and (f) we present the estimated posterior mean for the ROC curves over different ages.

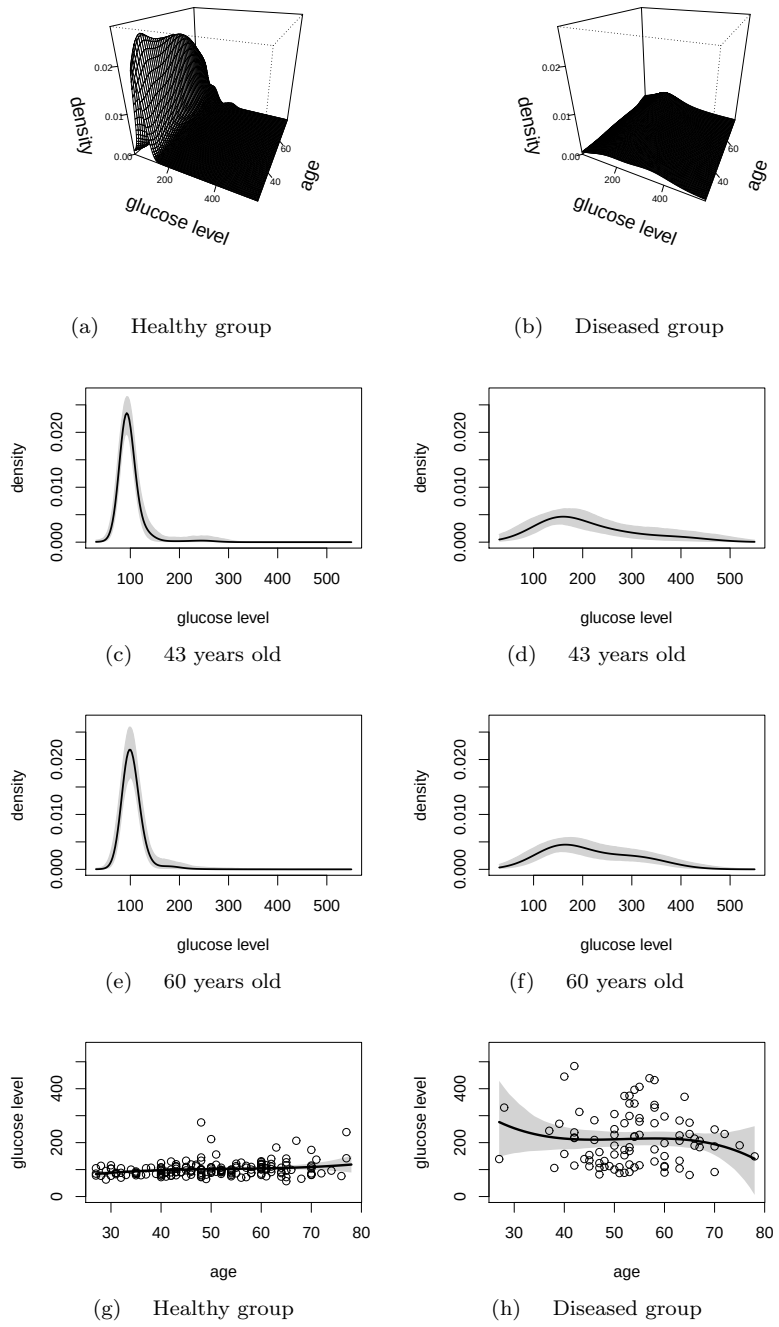
Specifically, we considered in Figure 8 (c–f) the ages 31, 43, 60, and 70, which correspond to the 5%, 25%, 75%, and 95% quantiles of the empirical distribution of the age, respectively. The corresponding AUC (95% point-wise HPD interval) were 0.909 $(0.661, 0.998)$, 0.877 $(0.752, 0.954)$, 0.887 $(0.788, 0.953)$, and 0.865 $(0.683, 0.964)$, respectively. Comparing these results with the one obtained by ignoring age, which was 0.885 $(0.823, 0.934)$, we see that ignoring age results in an over-or under-estimation of the AUC for certain ages. To examine the age effect further, Figure 8 (b) displays the posterior mean for the AUC as a function of the age. This figure clearly shows that age has an important impact on the discriminatory capacity of the glucose, with this marker having a better discriminatory capacity for ages between 27 and 70. After the age of 70 the discriminatory capacity of blood glucose as a marker to detect diabetes reduces substantially and, ignoring the factor age, will lead to an overestimated AUC for individuals older than 70 years old. We also point out that inferences are more precise for younger individuals than for older ones, where the credible band is wider.

4.3 Sensitivity analysis

We performed a sensitivity analysis using the B-splines DDP II mixture model (see, Section 3.4). Under this prior formulation, the log-pseudo-marginal likelihood (LPML) statistics for the diseased and healthy groups are -529.31 and -778.53, respectively. In turn, the LPML values obtained under the B-splines DDP mixture model (described in Section 2.3 and used in Section 4.2) are -528.78 and -774.88 for the diseased and healthy groups, respectively. Thus, from a predictive point of view, the latter model seems to be slightly preferable.

We point out that the two prior specifications are quite different. While the prior specification under the B-splines DDP mixture model is essentially data-driven, the prior specification under the B-splines DDP II mixture model is concentrated around the values listed in Section 3.4. In a sense, the latter prior is contradicting the data, since the glucose values vary from 57 to 484. However, the results under the two prior specifications, which are shown in Appendix E of the supplementary material, are not

contradictory.



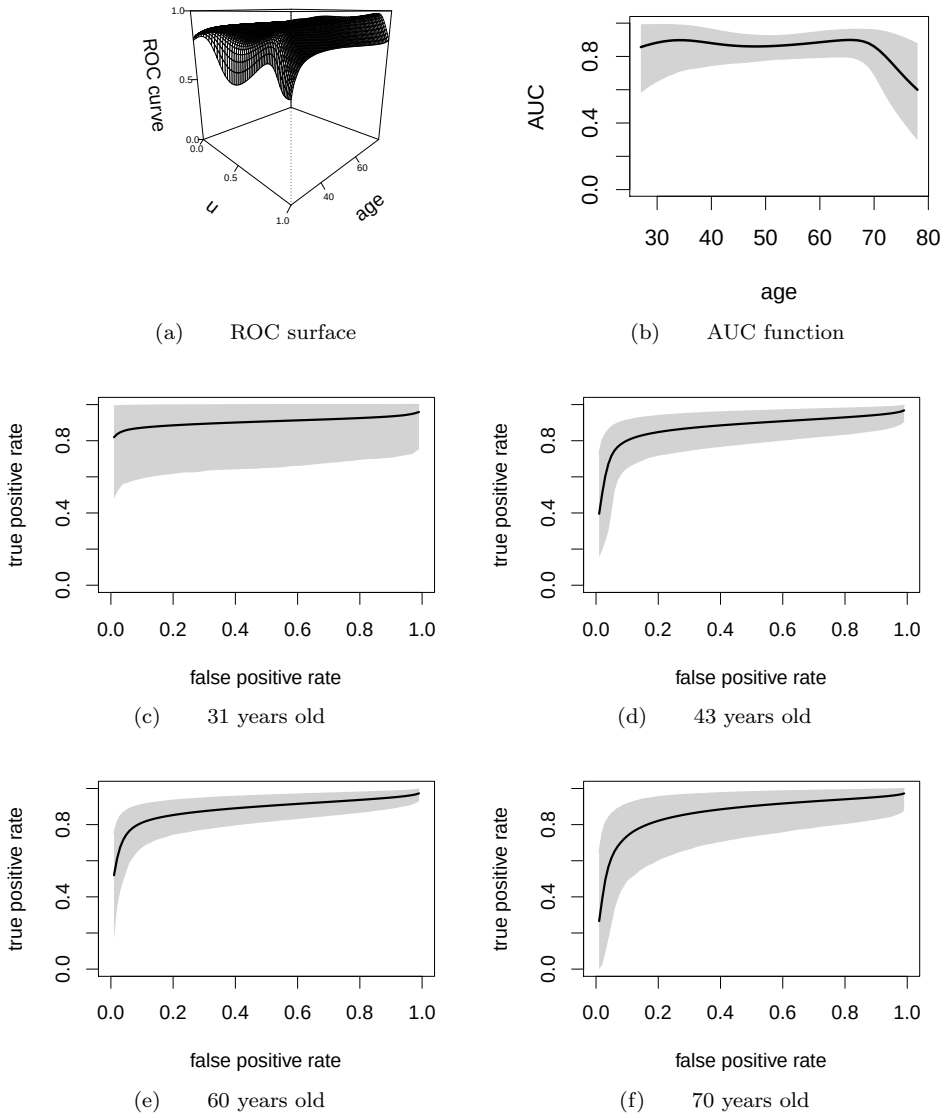
(a) Healthy group (b) Diseased group

(c) 43 years old (d) 43 years old

(e) 60 years old (f) 60 years old

(g) Healthy group (h) Diseased group

Figure 7: Glucose data: Conditional densities. Panels (a) and (b) display the surface of the posterior mean of the conditional densities across age for the healthy and diseased group, respectively. Panels (c) and (e), and (d) and (f) show the posterior mean and a 95% point-wise HPD band for the conditional densities corresponding to the 25% and 75% quantiles of the empirical distribution of the age in the healthy and diseased group, respectively. The posterior mean and 95% point-wise HPD band for the conditional mean function in the healthy and diseased group are displayed in panel (g) and (h), respectively.



(a) ROC surface (b) AUC function

(c) 31 years old (d) 43 years old

(e) 60 years old (f) 70 years old

Figure 8: Glucose data: Conditional ROC curve. Panel (a) displays the surface of the posterior mean of the conditional ROC curves across age. Panel (b) displays the posterior mean (solid line) and 95% point-wise HPD band for the area under the curve (AUC) as a function of the age. Panels (c), (d), (e) and (f) display the posterior mean and 95% point-wise HPD bands for the ROC curve corresponding to the 5%, 25%, 75% and 95% quantiles of the empirical distribution of the age, respectively.

5 Concluding remarks

We have proposed a Bayesian nonparametric framework for conditional ROC curve estimation using continuous and discrete predictors. Our approach is based on dependent Dirichlet processes and justified by the full support of the resulting model on the functional parameters of interest. Using simulated data, we have shown that an approximated version of the general model, based on B–splines, can outperform its competitors under non-standard assumptions for the 'true' underlying model. The results also suggest that there is little price to be paid for the extra generality when standard parametric assumptions hold.

Our methodology was applied to data concerning diagnosis of diabetes. We found that glucose has a good performance in diagnosing diabetes in young individuals, but its ability to distinguish diabetic and nondiabetic individuals decreases for older ages. This observation should be taken into account in the use of this marker in the clinical diagnosis of diabetes.

Acknowledgments

The first author thanks Antónia Amaral Turkman for support and Wenceslao González–Manteiga for having shared his expertise on kernel techniques with her. The research of V. Inácio de Carvalho is funded by the Portuguese Foundation for Science and Technology through PEst-OE/MAT/UI0006/2011 and PTDC/MAT/118335/2010. A. Jara's research is supported by Fondecyt grant 11100144. M. de Carvalho is funded by the Portuguese Foundation for Science and Technology through PEst-OE/MAT/UI0297/2011 and by the Fondecyt grant 11121186.

References

Alonzo, T. A. and Pepe, M. S. (2002). "Distribution-free ROC analysis using binary regression techniques." *Biostatistics*, 3: 421–432. 624

Barrientos, A. F., Jara, A., and Quintana, F. (2012). "On the support of MacEachern's dependent Dirichlet processes and extensions." *Bayesian Analysis*, 7: 277–310. 625

Blackwell, D. and MacQueen, J. (1973). "Ferguson distributions via Pólya urn schemes." *The Annals of Statistics*, 1: 353–355. 629

Branscum, A. J., Johnson, W. O., Hanson, T. E., and Gardner, I. A. (2008). "Bayesian semiparametric ROC curve estimation and disease diagnosis." *Statistics in Medicine*, 27: 2474–2496. 625

Cai, T. (2004). "Semiparametric ROC regression analysis with placement values." *Biostatistics*, 5: 45–60. 624

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2009). "Bayesian nonparametric non-proportional hazards survival modelling." *Biometrics*, 65: 762–771. 627, 628

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). "An ANOVA model for dependent random measures." *Journal of the American Statistical Association*, 99: 205–215. 627

De la Cruz, R., Quintana, F. A., and Müller, P. (2007). "Semiparametric Bayesian classification with longitudinal markers." *Journal of the Royal Statistical Society, Ser.* C, 56(2): 119–137. 637

Dubey, S. (1970). "Compound gamma, beta and F distributions." *Metrika*, 16: 27–31. 628

Eilers, P. H. C. and Marx, B. D. (1996). "Flexible smoothing with B-splines and penalties." *Statistical Science*, 11(2): 89–121. 627

Erkanli, A., Sung, M., Costello, E. J., and Angold, A. (2006). "Bayesian semiparametric ROC analysis." *Statistics in Medicine*, 25: 3905–3928. 625

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall. 631

Faraggi, D. (2003). "Adjusting receiver operating characteristic curves and related indices for covariates." *Journal of the Royal Statistical Society, Ser.* D, 52: 1152–1174. 624, 637

González-Manteiga, W., Pardo-Fernandéz, J. C., and Van Keilegom, I. (2011). "ROC curves in non-parametric location-scale regression models." *Scandinavian Journal of Statistics*, 38: 169–184. 624, 625, 631, 637

Hanson, T., Branscum, A., and Gardner, I. (2008a). "Multivariate mixtures of Polya trees for modelling ROC data." *Statistical Modelling*, 8: 81–96. 625

Hanson, T., Kottas, A., and Branscum, A. J. (2008b). "Modelling stochastic order in the analysis of receiver operating characteristic data: Bayesian non-parametric approaches." *Journal of the Royal Statistical Society, Ser.* C, 57: 207–225. 625

Hsieh, F. and Turnbull, B. (1996). "Nonparametric and semiparametric estimation of the receiver operating characteristic curve." *The Annals of Statistics*, 24: 24–40. 624

Inácio, V., Turkman, A. A., Nakas, C. T., and Alonzo, T. A. (2011). "Nonparametric Bayesian estimation of the three-way receiver operating characteristic surface." *Biometrical Journal*, 53: 1011–1024. 625

Jara, A. (2007). "Applied Bayesian non- and semi-parametric inference using DPpackage." *Rnews*, 7: 17–26. 630

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. L. (2011). "DPpackage: Bayesian semi- and sonparametric modeling in R." *Journal of Statistical Software*, 40: 1–30. 630, 637

Jara, A., Lesaffre, E., De Iorio, M., and Quintana, F. A. (2010). "Bayesian semiparametric inference for multivariate doubly-interval-censored data." *Annals of Applied Statistics*, 4: 2126–2149. 628, 637

Lloyd, C. J. (1998). "Using smooth receiver operating characteristic curves to summarize and compare diagnostic systems." *Journal of the American Statistical Association*, 93: 1356–1364. 624

MacEachern, S. N. (1994). "Estimating normal means with a conjugate style Dirichlet process prior." *Communications in Statistics: Simulation and Computation*, 23: 727–741. 629

— (2000). "Dependent Dirichlet processes." Technical report, Department of Statistics, The Ohio State University. 626

MacEachern, S. N. and Müller, P. (1998). "Estimating mixture of Dirichlet process models." *Journal of Computational and Graphical Statistics*, 7: 223–338. 629

Muliere, P. and Tardella, L. (1998). "Approximating distributions of random functionals of Ferguson-Dirichlet priors." *The Canadian Journal of Statistics*, 26: 283–297. 630

Neal, R. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9: 249–265. 629

Peng, L. and Zhou, X. H. (2004). "Local linear smoothing of receiver operating characteristic (ROC) curves." *Journal of Statistical Planning and Inference*, 118: 129–143. 624

Pepe, M. S. (1998). "Three approaches to regression analysis of receiver operating characteristic curves for continuous test results." *Biometrics*, 54: 124–135. 624, 625, 631

— (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford University Press. 624

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. 630

Richardson, S. and Green, P. J. (1997). "On Bayesian analysis of mixtures with an unknown number of components." *Journal of the Royal Statistical Society, Ser.* B, 59: 731–792. 628

Rodríguez-Álvarez, M. X., Roca-Pardiñas, J., and Cadarso-Suárez, C. (2011a). "ROC curve and covariates: extending the induced methodology to the non-parametric framework." *Statistics and Computing*, 21: 483–495. 624, 625, 631

Rodríguez-Álvarez, M. X., Tahoces, P. C., Cadarso-Suárez, C., and Lado, M. J. (2011b). "Comparative study of ROC regression techniques—applications for the computer-aided diagnostic system in breast cancer detection." *Computational Statistics and Data Analysis*, 55: 888–902. 624

Sarwar, N., Gao, P., Seshasai, S. R., Gobin, R., Kaptoge, S., Di Angelantonio, E., Ingelsson, E., Lawlor, D. A., Selvin, E., Stampfer, M., Stehouwer, C. D., Lewington, S., Pennells, L., Thompson, A., Sattar, N., White, I. R., Ray, K. K., and Danesh, J. (2010). "Diabetes mellitus fasting blood glucose concentration and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies." *The Lancet*, 375: 2215–2222. 637

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 2: 639–650. 626

Smith, P. J. and Thompson, T. J. (1996). "Correcting for confounding in analyzing receiver operating characteristic curves." *Biometrical Journal*, 7: 857–863. 637, 639

Wild., S., Roghic, G., Green, A., Sicree, R., and King, H. (2004). "Global prevalence of diabetes: estimates for 2000 and projection for 2030." *Diabetes Care*, 27: 1047–1053. 637

Xu, L., Hanson, T., Bedrick, E., and Restrepo, C. (2010). "Hypothesis tests on mixture model components with applications in ecology and agriculture." *Journal of Agricultural, Biological, and Environmental Statistics*, 15: 308–326. 628

Zhou, X. H. and Harezlak, J. (2002). "Comparison of bandwidth selection methods for kernel smoothing of ROC curves." *Statistics in Medicine*, 21: 2045–2055. 624

Zou, K. H., Hall, W. J., and Shapiro, D. E. (1997). "Smooth nonparametric receiver operating characteristic (ROC) curves for continuous diagnostic tests." *Statistics in Medicine*, 16: 2143–2156. 624